Decision Support System on Prediction of Heart Disease Using Data Mining Techniques

Ms. Shinde Swati B.
PG Student, Department of Information Technology, DKGOI-FOE Swami-Chincholi,Pune, Maharashtra, India.
s.swatishinde@gmail.com

Prof. Amrit Priyadarshi
Assistant Professor, Department of Computer Engg., DKGOI-FOE Swami-Chincholi, Maharashtra, India.
amritpriyadarshi@gmail.com

Abstract - Hospital Industry collects large amount of data, which is not perfect to diagnose the disease. Mining of huge data must be necessary ,to diagnose the disease. Medical industry has rich information but poor knowledge. Mining of data is necessary for discovering the hidden patterns and relationships necessary for decision making. Data Mining Techniques are used in medical decision support system for prediction of various diseases. Prediction or Diagnosis of Heart Disease is one of the most important application of Data Mining Techniques.

This paper describes the diagnosis of heart disease using data mining techniques such as Naive Bayesian and K-Nearest Neighbors (KNN). Traditional support system can answer only simple queries but can not answer complex queries like 'What If'. This system can answer complex queries like what if queries. Using medical attributes such as sex, age, blood sugar, blood pressure etc. can predict the likelihood of Patient having heart disease. Data mining extract the knowledge such as hidden patterns and relationship between medical attributes related to heart disease . This training tool is useful for medical students and nurses for prediction of heart disease.

Keywords- Heart Disease, Data Mining, Decision Support System , Naive Bayesian , KNN, Hospital Data, Diagnosis System etc.

I. INTRODUCTION

Data mining is defined as finding hidden patterns and relationships in a database. it is also called as exploratory data analysis, data driven discovery. Traditional queries access database using a Structured Query Language named SQL [1]. Output of the query consist data from the database that satisfies the query. The output is a subset of the database but it may contain aggregations. Data mining access the database which is different from the traditional access. The data must be cleaned and modified to better support to the mining process. The output of the data mining query is not a subset of the database, instead it is the output of some analysis of the contents of database.

Data mining is used to extract useful information from huge amount of data. It includes various types of areas like machine learning, statistics, pattern recognition, artificial intelligence and data visualization [2]. Data Mining has different types of models such as predictive and descriptive model. Predictive Model makes a prediction about values of data using known results found from different data. Descriptive model identifies patterns and relationships in data. It explore the properties of data , not to predict new properties. The Predictive models such as Naive Bayesian and KNN are used for the prediction of heart disease.

The huge amount of data from the hospital contains charts, image, text and numbers. Rarely these data are used for clinical decision making. It contain hidden information which can be used for prediction of several diseases. Some hospitals use traditional decision support system which can answer only simple queries but can not answer complex queries. They can answer simple queries such as "How many patients are there having heart disease between the age 30 to 50?". But, they can not answer complex queries like "Predict the probability of patient getting the heart disease from the given record." [3]. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [4].

II. LITERATURE SURVEY

Intelligent Heart Disease Prediction System uses Decision Tree, naive Bayesian and Neural Network was proposed by Sellappan Palaniappan et ul, Rafiah Awang [4]. This system answers the complex queries which can not be answered by conventional system. The Diagnosis of Heart Disease using Neural Network was proposed by Niti Guru et al [5]. This system uses sample database of patients' records. In this system input is tested and uses 13 trained attributes such as sex, age, blood pressure etc. When unknown input is given then system compares this input with trained data and produce the list of probable disease.

Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm proposed by m. Anbarasi, e. Anupriya and n.ch.s.n.Iyengar. This system predict the heart disease by reducing the input attributes such as Chest pain type, Resting blood pressure, Exercise Induced angina, old peak, No. of vessels colored, Maximum heart rate achieved. Milan Kumari et al. [6], proposed a system which contain data mining classification techniques such as Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM). These technique are used for cardiovascular disease dataset. Accuracy of Decision Tree, ANN and SVM are 79.05%, 80.06% and 84.12% respectively. The extraction of patterns from the heart disease warehouse for prediction was proposed by Shantakumar B.Patil et al [7]. First, the data warehouse is pre-processed then the heart disease warehouse is clustered with the K-means clustering algorithm. The patterns which are extracted are mined with the MAFIA algorithm. In addition, the patterns of heart attack prediction are selected. The neural network is trained with the selected patterns for the prediction of heart disease.

III. RESEARCH OBJECTIVE

The main objective of this paper is to predict or diagnose the heart disease using data mining techniques such as Naive Bayesian and KNN algorithm etc. This system generates and extracts hidden information i.e. patterns and relationships between different attributes from the historical heart related database. It can answer complex queries which can not be answered by traditional decision support system. If this system is used, treatment cost may be reduced which is affordable to each and every patient. It also improves the quality of clinical decisions.

IV. DATA SOURCE

A total of 1000 records with 13 attributes were obtained from the database. These records are divided into two dataset i.e training dataset (700) and testing dataset (300). Records for each set are selected randomly to avoid bias.
The ‘Diagnosis’ attribute is used to predict the heart disease with value “2” for patient having heart disease and “1” for patient having no heart disease. The ‘PatientID’ attribute is used as key and others are input attributes.

Predictable attributes

1. Diagnosis (value 2 – Patient having heart disease and value 1- Patient having no heart disease)

Key Attribute

1. PatientID – Patient’s identification number

Input Attributes

1. Age (value 1:<=40, value 2:<=60 and >40, value 3: >60)
2. Sex (value 0: female, value 1: male)
3. Chest Pain Type (value 1:Low, value 2: Medium, Value 3: High, Value 4: Very High)
4. Blood Pressure (value 1:<=80, value 2:<=120 and >80, value 3: >120)
5. Blood Sugar (value 0: Low, value 1: High)
6. Serum Cholesterol(value 1:<=180, value 2:<=400 and >180, value 3: >400)
7. Resting ECG (value 0: normal, value 1: wave abnormality, value 2: showing probable or definite left, ventricular hypertrophy)
8. Heart Rate (value 1:<=120, value 2:<=180 and >120, value 3: >180)
9. Exercise Induced Angina (value 0: Low, value 1: High)
10. Oldpeak (ST depression value 1:<=1, value 2:<=2.5 and >1, value 3: >2.5)
11. Slope of the peak Exercise (value 1: unslopping, value 2: flat, value 3: downslopping)
12. No.of major vessels(value 1:Low, value 2: Medium, Value 3: High, Value 4: Very High)
13. Thal (value 3: normal; value 6: fixed defect; value 7:reversible defect)

Figure 1. Description of Attributes

V. DATA MINING TECHNIQUES

Fayyad defines data mining as “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database”. It is an exploratory data analysis, trying to discover useful patterns in data. This system is implemented using Naive Bayes or Bayes’ rule and K Nearest Neighbors (KNN) algorithm.

1. *Naive Bayes algorithm*

Bayes’ rule is the basis for many machine learning and data mining methods [8]. Naive bayes algorithm has been proposed that is based on bayes rule of conditional probability. By analyzing the contribution of each “independent” attribute, a conditional probability is determined. The approach is called “naive” because it assumes the independence between the various attribute values. This algorithm is used to create models with predictive capabilities. Bayes rule is a technique to estimate the likelihood of a property from

the given set of data. Naïve Bayes classification can be viewed as both descriptive and predictive type of algorithm. The probabilities are descriptive and used to predict the class membership for a target tuple
This approach can easily handle missing values by simply ommitting that probability when calculating the likelihoods of membership in each class.

1) Algorithm

Given the Hospital data set

1. Estimate the prior probability P (cj) for each class by counting how often each class occurs in the training data.
2. For each attribute Xi find P (xi) by counting the number of occurrences of each attribute value.
3. Find probability P (xi/cj) by counting how often each value occurs in the class in the training data.
4. Do this for all attributes and all values of these attributes. To classify a target tuple estimate P (ti/cj) = ∏Pk=1 P (xik/cj).
5. Calculate P (ti). This can be done by finding the likelihood that this tuple is in each class and then adding all this values.
6. Find posterior probability P (cj/ti) for each class. It is the product of conditional probabilities for each attribute value.
7. Select class with highest probability value of P (cj/ti) value for test tuple

2) Mathematical Formulae

P (HeartDis Yes) = No.of Records with Result
Yes / Total no. of Records

P (HeartDis N) = No.of Records with Result
No / Total no. of Records

P (t/yes) = P (Age (low) yes) * P (Sex (Male) yes) * P (BP (High) yes) * P (Chol (High) yes) * P (Heart_Rate (High)yes)*P(Vessels(High)yes)*P(Chest_Pain(High)yes)*P(ECG(High)yes)*P(Exer_angina(High)yes)*P(old_peak(High)yes)*P(Thal(High)yes)*P(Blood_sugar(High)yes) * P (Slope_peak(High)yes)

P (t/no) = P (Age (low) no) * P (Sex (Male) no) * P (BP (High) no)* P (Chol (High) no) * P (Heart_Rate (High) no)*P(Vessels(High)no)*P(Chest_Pain(High)no)*P(ECG(High)no)* P(Exer_angina(High) no)* P (old_peak(High) no)* P (Thal(High) no)* P (Blood_sugar(High)
no)* P (Slope_peak(High) no)

P (Likelihood of yes) = P (t/yes) * P (Heart_Disease yes)

P(Likelihood of no)= P (t/no) * P (Heart_Disease no)

Now we find the total probability,
P(yes/t) = P (t/yes) * P (Heart_Disease yes) **/** P (T)
P(no/t) = P (t/no) * P (Heart_Disease no) / P (T)

If P (yes/t) >= P (no/t) then input query is classified as Heart Disease category
Else No Heart Disease category

B. K Nearest Neighbors Algorithm

One common classification scheme based on the use of distance measures is that of K Nearest Neighbors (KNN). K-NN is a type of distance-based learning. The k-nearest neighbor algorithm is amongest the simplest of all machine learning algorithms. KNN is a *non-parametric lazy learning* algorithm. It means that it does not make any assumptions on the underlying data distribution. This is very useful, as in the real world most of the practical data does not obey the typical theoretical assumptions . The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. When a classification is to be made for a new item, its distance to each item in the training set must be determined. Only the K closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from this set of K closest items.

1) Algorithm

Given the Hospital Data Set

1. Enter the value of K

2. Give the Input Query
3. Find the Square Distance to Query Instance
4. Map Square Distance w.r.t. Output and Store it in Array
5. Sort the Square Distance and also sort the Output w.r.t. Square Distance
6. Find out the Count of Yes_Heart_Disease and No_ Heart_Disease up to the K
7. Compare Yes_Heart_Disease Count and the No_Heart_Disease Count and
8. If(Count(Yes_Heart_Disease) >= Count (No_Heart_Disease))
9. Print ("Record is classified as the Heart_Disease Category");
10. Yes_Count++; Else
    Print ("Record is classified as the No Heart_Disease Category");
11. No_Count++;
12. Calculate Accuracy

2) Mathematical Formulae

Euclidian Square Distance Formulae

Square Distance = $(x_{age} - x_{in\_age})^2 + (x_{BP} - x_{in\_BP})^2 + (x_{chol} - x_{in\_chol})^2 + (x_{hrtRate} - x_{in\_hrtRate})^2 + (x_{oldPeak} - x_{in\_oldPeak})^2 + (x_{SlopePeak} - X_{in\_SlopePeak})^2 + (x_{vessels} - X_{in\_vessels})^2 + (x_{thal} - x_{in\_thal})^2 + (x_{chestPain} - x_{in\_ChestPain})^2 + (x_{oldPeak} - x_{in\_oldPeak})^2 + (x_{angina} - x_{in\_angina})^2 + (x_{ecg} - x_{in\_ecg})^2 + (x_{BloodSugar} - x_{in\_BloodSugar})^2$

Accuracy Calculation

Accuracy refers to the percentage of correct predictions made by the model compared with actual classifications in the test data.

Accuracy = Total no. of Correctly Predicted Record
/ Total no. of training Record

VI. PROPOSED SYSTEM

The main goal of this system is to predict heart disease using different data mining techniques. Raw hospital data set is used and then preprocessed and transformed the hospital data set. Then apply the data mining techniques i.e. K-NN and Naïve Bayes algorithm on the transformed data set. After applying the data mining algorithm , heart disease is predicted and then accuracy is calculated.

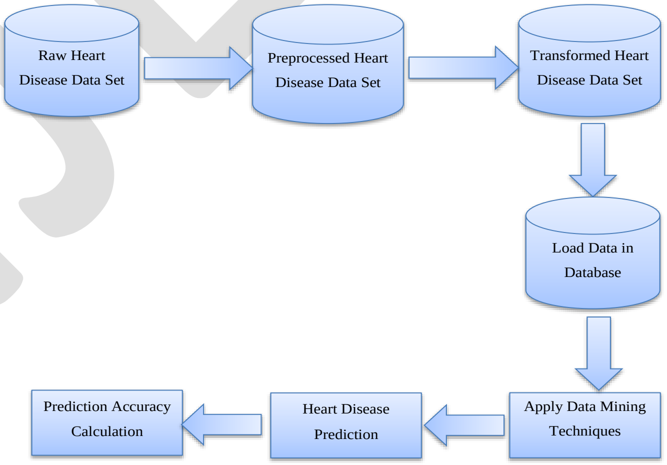
Raw Heart Disease Data Set
Preprocessed Heart Disease Data Set
Transformed Heart Disease Data Set
Load Data in Database
Apply Data Mining Techniques
Heart Disease Prediction
Prediction Accuracy Calculation

Figure 2: System Architecture

VII. PRACTICAL RESULTS AND ENVIRONMENT

In this section , practical environment as input of the system, operating environment, software and hardware requirement and the result of the system are represented.

A. Input Database

Total 13 attributes i.e. input attributes such
as age, sex, blood sugar, cholestrol, heart rate, chest pain etc. are provided to the system.

B. Hardware Requirements

1. Processor : Intel Dual-Core processor.
2. RAM : 512 MB.
3. HDD : 10 GB.

C. Software Requirements

1. Operating System - Windows 2000,Windows 2007/XP.
2. Documentation -MS Word, MS PowerPoint, MS Excel.
3. Database - Oracle 10g.
4. Language - Java
5. Software Tools - Eclipse IDE, SQL Loader, Java J2SE
   JDK 1.6

D. Result of the System

Table 1: Accuracy Of Algorithm

Algorithm used	**Accuracy**
Naive Bayes	84%
KNN	76%

From the result , it is concluded that Bayesian algorithm works better than KNN algorithm.

VIII. CONCLUSION

This system uses two data mining techniques such as Naive Bayesian and K- Nearest Neighbor algorithm. This system extracts patterns and relationships from the historical database. This system is useful in hospital for prediction of disease. After execution, it found that Naive Bayesian works better than KNN algorithm. You can add more attributes to enhance and expand the system. You can also use other data data mining techniques such as clustering, time series, association rules etc. We can also use Text Mining for mining the data which is not structured. We can also integrate data mining and text mining.

REFERENCES:

[1] Margaret H. Dunham, Southern Methodist University ,'Data Mining- Introductory and Advanced Topics, ISBN: 0130888923 published by Pearson Education, Inc.,Sixh Impression (2009).

[2] Han and kamber, "Data mining concepts and techniques", 2nd edition(2010)

[3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[4] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management.

[5] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi ,Business Review, Vol. 8, No. I (January - June 2007).

[6] Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol.

2, Iss ue 2, June 2011.

[7] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research, ISSN 1450-216X, Vol.31 No.4 (2009), pp.642-656.

[8] Tang, Z. H., MacLennan, J.: “Data Mining with SQL Server 2005”, Indianapolis: Wiley, 2005